An Experimental Study of Truth-Telling in a Sender-Receiver Game*

5 September 2006

Santiago Sánchez-Pagés

Corresponding Author. Edinburgh School of Economics, University of Edinburgh, 50 George Square, EH8 9JY, Edinburgh, U.K. Tel.: +44 (0) 131 651 3005, Fax: +44 (0) 131 650 4514, E-Mail: ssanchez@staffmail.ed.ac.uk

Marc Vorsatz

Department of Economics (AE1), Maastricht University, P.O. Box 616, 6200MD Maastricht, The Netherlands. E-Mail: m.vorsatz@algec.unimaas.nl

Abstract

A recent experimental study of Cai and Wang (2005) on strategic information transmission reveals that subjects tend to transmit more information than predicted by the standard equilibrium analysis. To evidence that this overcommunication phenomenon can be explained in terms of a tension between normative social behavior and incentives for lying, we show in a simple sender-receiver game that subjects incurring in costs to punish liars tell the truth more often than predicted by the logit agent quantal response equilibria whereas subjects that do not punish liars after receiving a deceptive message play, on the aggregate, equilibrium strategies. Thus, we can partition the subject pool into two groups, one group of subjects with preferences for truth-telling and one taking into account only material incentives.

Keywords: Morally Consistent Behavior, Procedural Justice, Strategic Information Transmission, Truth-Telling.

JEL-Numbers: C72, C73, D83.

*We thank Jordi Brandts for his support. We are grateful to Raúl López, Rosemarie Nagel, Jordi Massó, Alvin Roth, and Larry Samuelson for their useful comments. The questions raised by the associate editor and two anonymous referees helped us to improve the paper significantly. We also thank Marco Faravelli for helping us to conduct the first series of experiments. The usual disclaimer applies. This research has been possible thanks to the financial support of the Development Research Trust Fund of the University of Edinburgh, the small scale research project 05/05 of Maastricht University and the small research grants scheme of the Royal Economic Society. Vorsatz acknowledges financial support from the fellowship 2001FI 00451 of the Generalitat de Catalunya and the research grant BEC2002-02130 of the Ministerio de Ciencia y Tecnología de España.

1 Introduction

Individuals who lie about their private information can obtain a higher payoff at the costs of others in several situations.[1] But by behaving strategically individuals disrespect one of the oldest ethical principles, a social norm telling us *not to lie*. This tension between incentives and normative social behavior makes it difficult to predict the outcome of this type of interactions. It is our objective to show, with the help of an experiment, that in situations that can be modeled as a particularly simple *sender-receiver game*, a considerable number of subjects have preferences for truth-telling, whereas the rest of the subjects follow only material incentives.

Strategic information transmission, introduced by Crawford and Sobel (1982), is an obvious way of modeling the tension described above. In this class of games, the *sender* has private information about the true state of the world. She transmits a message about the actual state to the *receiver* who takes a subsequent action that is payoff-relevant for both participants. The main insight of Crawford and Sobel (1982) is that less information about the true state is transmitted as the preferences of the sender and the receiver become less aligned.

In the first experimental study on strategic information transmission, Dickhaut et. al (1995) corroborated this theoretical prediction. More recently, Gneezy (2005) has shown that if preferences are conflictive (whenever an outcome is good for the receiver it is bad for the sender and vice versa), then the probability of lying is increasing in the potential gains to the sender and decreasing in the potential loss to the receiver. Finally, Cai and Wang (2005) have offered clear experimental evidence of an *overcommunication phenomenon*: Senders truthfully reveal more private information than predicted by the most informative equilibrium of the standard model of preference maximization. Although the authors explain this abnormality successfully by means of a behavioral type analysis (see among others Nagel (1995), Costa-Gomes et. al (2001) and Crawford (2003)) and the quantal response equilibrium concept (McKelvey and Palfrey (1995) and (1998)), they leave it as an open question whether the overcommunication phenomenon is caused by social preferences such as trust or honesty.

Our aim is to show that the tension between incentives and normative social behavior is the driving force underlying the overcommunication result. To this end, we study the experimental behavior of a group of subjects in two very

[1] Examples include income tax evasion (Alingham and Sandmo (1972)), oligopolistic competition (Galor (1986)), financial advice (Morgan and Stocken (2003)), and electoral competition (Heidhues and Lagerlof (2003)).

similar constant-sum sender-receiver games. The *Benchmark Game* proceeds as follows: In the beginning of the game, one out of two payoff tables is randomly chosen. The selected table determines players' (strictly positive) payoffs as a function of the receiver's action to be taken later on. Then, the sender, who is the only player informed about Nature's choice, submits a message about the actual payoff table. Hence, she implicitly decides whether to tell the truth or to lie. After observing this message, the receiver takes an action that reveals whether he trusted or distrusted the sender. Finally, both players are paid accordingly.

Since the payoff tables are constructed in such a way that the preferences of the sender and the receiver are completely opposed, the sender does not have an incentive to transmit any information, or, to say it differently, the sender plays a strategy such that the posterior beliefs of the receiver remain equal to the prior beliefs. Given our model specification, only those strategies in which the sender lies with probability one-half generate these beliefs consistently and can thus be supported in equilibrium. This action is foreseen correctly by the receiver, and, as a consequence, random play for both individuals is the only sequential equilibrium of the Benchmark Game (Proposition 1).

Bounded rationality is one major reason why subjects fail to play equilibrium strategies in laboratory experiments. One model of bounded rationality that has been successfully employed to explain experimental data is the agent quantal response equilibrium (AQRE) of McKelvey and Palfrey (1998). Here we solve for the logit-AQRE, which is parameterized by $\lambda \in [0, \infty)$. Since it is well known that random play is the unique logit-AQRE for all sequential games if $\lambda = 0$ and every logit-AQRE is also sequential equilibrium when λ tends to infinity, it is not surprising that random play is the unique logit-AQRE of the Benchmark Game (Proposition 2).

In the first step of our experimental analysis we show that subjects playing the Benchmark Game in the role of the sender lie significantly less than predicted by equilibrium theory (Hypothesis 1). Then, in order to provide evidence that this result is caused by a considerable number of subjects with preferences for truth-telling we extend our original set-up. In the *Punishment Game*, the receiver is informed about the actual payoff table once he has taken an action. Finally, he chooses between accepting the payoff distribution induced by the Benchmark Game and reducing the payoffs of both players to zero.

Whereas rational individuals should never punish the sender (this action is costly), receivers may do so according to the logit-AQRE concept because it is assumed that individuals make mistakes when they try to maximize their payoffs. Yet, the punishment rate must depend only on the payoff the receiver

foregoes (Proposition 4). On the other hand, the inequity aversion models of Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) are also able to explain the empirical observation that some individuals are willing to pay money in order to reduce income disparities. Hence, in these models the punishment rate depends upon the whole payoff distribution. Finally, Brandts and Charness (2003) have recently uncovered an even more complex type of preferences: Individuals seem not only to take into account the whole payoff distribution, rather the notion of *procedural justice*[2]-the utility attached to a payoff distribution depends on how this distribution has been reached- plays a crucial role in socio-economic interactions. The authors study in the laboratory a game in which the sender transmits a message regarding her/his intended play in a 2×2 simultaneous move game and show that the receiver's willingness to punish the sender after revealing the result from the simultaneous move game depends on whether or not the sender played according to the reported message.

We base our predictions for the Punishment Game on the prevalence of the latter type of preferences and we are able to reject the other two explanations by looking at the punishment rates after different histories. Since the game is symmetric, histories can be summarized by whether or not a message is truthful and whether or not the receiver trusts that message. The punishment rates are equal to 0% after history (truth,trust), 1.6% after (lie,distrust), 5.4% after (truth,distrust), and 25.2% after (lie,trust). Since the payoff distribution after history (truth,distrust) is equal to the one after (lie,trust), we take the difference in the punishment rates after these histories (lie,trust) and (truth,distrust) as an estimator for the importance of procedural justice. We confirm that this difference is significantly greater than zero (Hypothesis 2).

In the next step of our analysis we use the previous results to show that the excessive truth-telling in the Benchmark Game can be explained in terms of social preferences for truth-telling or normative social behavior. First, we use the punishment treatment to identify all subjects with strong concerns for procedural justice. In particular, we find that 15 out of 66 individuals punish liars frequently after history (lie,trust). Not surprisingly this group of individuals punishes the sender in over 81% of all observations after this history and accounts for 90% of all punishments related to procedural justice. Then, we study how these subjects behave in the role of the sender. It turns out that they tell the truth in 70.6% of the occasions whereas the rest of subjects do so only in 52% (the overall percentage of truth-telling in the Punishment Game is equal to 56.3% and not significantly different from the corresponding

[2]The concept of procedural justice has been introduced in decision theory by Sen (1997) as an extension of the standard model of preference maximization over material outcomes.

percentage in the Benchmark Game). This result supports our conjecture that individuals with a strong sense for procedural justice should, consistently, be responsible for the excessive truth-telling (Hypothesis 3). We refer to this as *morally consistent behavior*.

Finally, we ask how robust our results are with respect to an increase in the inequality of the potential payoff distribution while maintaining incentives the same. In the particular case we study, the sender can now gain more by deceiving the receiver if the latter trusts the sender's message and punishments hurt the sender more without increasing the associated cost to the receiver. We observe two main changes (Hypothesis 4): (1) The excessive truth-telling in the Benchmark Game vanishes, because the senders tell the truth only in 50.6% of all observations. This finding is consistent with the study of Gneezy (2005) who observed that the probability of lying is increasing in the potential gains to the sender. (2) The percentage of punishments after history (lie,trust) increases from 25% to over 42%. Moreover, since the punishment rate after history (truth,distrust) raises "only" from 5% to 13%, we conclude that procedural justice plays now an even more central role. All other results are very robust. Most importantly, subjects who punish liars frequently after history (lie,trust) are again responsible for the excessive truth-telling in the Punishment Game (the overall percentage of truth-telling in this game is now equal to 56.4%). These subjects tell the truth in 69.8% of all occasions whereas the rest do so only in 51.1%.

The remainder of the paper is organized as follows: In the next Section, we formally introduce the games and our experimental hypotheses. In Section 3, we explain the experimental procedures. Afterwards, we present our results and perform the robustness analysis. We conclude in Section 5. The proofs of the Propositions and the instructions of the Punishment Game are relegated to the Appendix.

2 Theoretical Analysis and Predictions

In this Section, we introduce the Benchmark and the Punishment Game and derive several null hypotheses from the corresponding quantal response equilibria. Moreover, we present our alternative hypotheses deduced from the incorporation of preferences for truth-telling.

The Benchmark Game

Let $N = \{\text{sender,receiver}\}$ be the set of players. At the beginning of the game, Nature picks payoff table A and B with equal probability, e.g. $p(A) =$

$p(B) = 0.5$. Only the sender is informed about the payoff table actually chosen. Selecting table $\theta \in \Theta = \{A, B\}$ means that final payoffs are realized according to θ. Both tables depend only on the action U or D taken by the receiver later on. We assume in both tables that $x > 1$.

Put Table 1 about here (caption: Payoff Tables)

After the sender has been informed, she chooses a mixed strategy with support on the message space $M = \{A, B\}$. Formally, if Nature selects table A, the sender communicates with probability $p(A|A) \equiv p_A$ that table A represents the actual payoff scheme. Thus, she lies in this case with probability $p(B|A) = 1 - p_A$. Similarly, if Nature selects table B, she communicates with probability $p(B|B) \equiv 1 - p_B$ that table B represents the actual payoff scheme. Thus, she lies in this case with probability $p(A|B) = p_B$.

Next, we describe the receiver's belief system. If $m = \{A\}$ (the sender transmits message A), the receiver believes with probability $\mu(A|A) \equiv \mu_A$ that the actual payoff scheme is represented by table A whereas he thinks with probability $\mu(B|A) = 1 - \mu_A$ that table B is the one determining payoffs. If $m = \{B\}$, the receiver believes with probability $\mu(A|B) \equiv \mu_B$ that table A determines payoffs and with probability $\mu(B|B) = 1 - \mu_B$ that table B is the one doing so. Taking into account these beliefs, the receiver chooses a mixed strategy with support on the action set $\mathcal{A} = \{U, D\}$. Formally, if $m = \{A\}$, the receiver takes action U with probability $q(U|A) \equiv q_A$ and action D with probability $q(D|A) = 1 - q_A$. Similarly, if $m = \{B\}$, the receiver takes action U with probability $q(U|B) \equiv q_B$ and action D with probability $q(D|B) = 1 - q_B$. Finally, both individuals receive their payoff. We denote by $u(a, \theta)$ and $v(a, \theta)$ the payoff of the sender and the receiver when the receiver takes action $a \in \mathcal{A}$ and the true state is $\theta \in \Theta$. ■

Put Figure 1 about here (caption: The Benchmark Game). The corresponding graphic file is called figure2.eps and it should be scaled down to about 50%.

The Benchmark Game is well suited to analyze the tension between social preferences for truth-telling and material incentives for two reasons. First, in order to minimize the possibility of mistakes made by subjects, the Benchmark Game has a simple payoff structure and a very intuitive set of equilibria. Second, truth-telling is a dichotomous choice in the sense that (a) there are only two state variables and (b) the sender's strategy set boils down to the messages *truth* and *lie*. This is important, because otherwise a message may contain a richer meaning. To see this suppose that the state and message space are

both equal to $\{1,2,3\}$. In this case individuals do not only tell the truth or lie, because they also choose a "level" of deceit whenever the true state is $\{1\}$ or $\{3\}$. Hence, richer state and message spaces give room to a wide variety of behaviors and to a complexity that lies out of the scope of the paper.[3] According to Proposition 1 the set of sequential equilibria of the Benchmark Game does not depend on x. Yet, since different values of x give raise to more or less fair payoff distributions, we are likely to observe an effect in our experimental results. To take this into account we set x equal to 2 in our first experimental series and perform afterwards a robustness analysis by considering the case when x is equal to 9.[4]

Proposition 1 *The set of sequential equilibria of the Benchmark Game is given by the set of strategies* $(p_A^*, p_B^*, q_A^*, q_B^*) = (p, p, q, q)$*, where* $p, q \in [0,1]$*, and the supporting belief system* $(\mu_A^*, \mu_B^*) = \left(\frac{1}{2}, \frac{1}{2}\right)$*.*

Proof: See Appendix A. □

The intuition of Proposition 1 is as follows: Since preferences are not aligned, the sender plays a strategy that leaves the receiver's prior beliefs unchanged. The strategies generating these posterior beliefs in a consistent manner are all those in which the sender submits message A with a constant probability, e.g. $p_A = p_B = p \in [0,1]$. To see this note that if the sender plays for example the strategy "always transmit message A" (this strategy is equal to $p_A = p_B = 1$), then the receiver does not get any additional information from the message. Hence, the receiver can as well ignore it. This game becomes thus equivalent to the following one: Nature selects the tables A and B with equal probability and the receiver chooses q (the probability to play U) to maximize his expected payoff. Since the expected payoff is equal to $p(A)(q+x(1-q))+p(B)(xq+1-q) = \frac{1+x}{2}$ and thus independent of q, any constant strategy $q_A = q_B = q \in [0,1]$ is optimal.

It is well known that experimental subjects often fail to play equilibrium strategies. One model of bounded rationality that has proved to be successful in explaining these experimental results is the Agent Quantal Response Equilibrium (AQRE) of McKelvey and Palfrey (1995). Its central idea is that individuals make mistakes when they try to maximize their payoffs but have correct beliefs about the opponents actions. In the logit-AQRE, which is parameterized by $\lambda \in [0,\infty)$, the sender transmits message m in state θ with probability

[3] The importance of the size of the message space is reported by Blume et. al (1998). The authors show that in a sender-receiver game with multiple equilibria it depends on the size of message space whether subjects converge to play a separating or a pooling equilibrium.

[4] We thank an anonymous referee for making this important suggestion.

$\tilde{p}(m|\theta) = \frac{e^{\lambda u(m|\theta)}}{\sum_{i \in M} e^{\lambda u(i|\theta)}}$, where $u(i|\theta) = \sum_{a \in \mathcal{A}} \tilde{q}(a|i) \cdot u(a, \theta)$ denotes the expected payoff from sending message i in state θ. Similarly, the receiver chooses action a after observing message m with probability $\tilde{q}(a|m) = \frac{e^{\lambda v(a|m)}}{\sum_{j \in A} e^{\lambda v(j|m)}}$. Here, $v(j|m)$ corresponds to the expected payoff of taking action j upon message m; that is, $v(j|m) = \sum_{\theta \in \Theta} \tilde{\mu}(\theta|m) \cdot v(j, \theta)$, where $\tilde{\mu}(\theta|m) = \frac{\tilde{p}(m|\theta)}{\sum_{i \in \Theta} \tilde{p}(i|\theta)}$ is the receiver's posterior belief about the state θ when he observes message m.

For most sequential games it is impossible to find a closed-form solution for the set of logit-AQRE, but here this task turns out to be rather simple. The intuition for this runs as follows: It is well known that random play is the only logit-AQRE for all sequential games when $\lambda = 0$ and that any logit-AQRE is also a sequential equilibrium of the underlying game when λ tends to infinity (see Theorem 2 in McKelvey and Palfrey (1998)). Since random play is the only sequential equilibrium of the Benchmark Game according to Proposition 1, random play has also to be the only logit-AQRE for all λ. Thus, the only remaining question is which sequential equilibrium is selected by the unique logit-AQRE. It turns out that it is the symmetric one.

Proposition 2 *The unique logit-AQRE of the Benchmark Game is given by the set of strategies* $(\tilde{p}_A^*, \tilde{p}_B^*, \tilde{q}_A^*, \tilde{q}_B^*) = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$ *and the supporting belief system* $(\tilde{\mu}_A^*, \tilde{\mu}_B^*) = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Proof: See Appendix B. □

We can derive from Proposition 2 the equilibrium levels of truth-telling and trust. Given p_A and p_B, the probability that the sender lies in the Benchmark Game is equal to $l_b(p_A, p_B) = p(A)(1 - p_A) + p(B)p_B$. With a slight abuse of notation let $p(m = i)$ be the probability that the sender transmits message $i \in \{A, B\}$. Then, the probability that the receiver trusts the sender in the Benchmark Game is equal to $t_b(p_A, p_B) = p(m = A)\mu_A + p(m = B)(1 - \mu_B)$. Proposition 3 establishes that the sender lies with probability one-half in all sequential equilibria of the Benchmark Game, a strategy foreseen correctly by the receiver in terms of trust.

Proposition 3 *Let* (p_A^*, p_B^*) *be a sequential equilibrium strategy for the sender in the Benchmark Game. Then,* $l_b(p_A^*, p_B^*) = t_b(p_A^*, p_B^*) = \frac{1}{2}$.

Proof: The proof is straightforward and thus omitted. □

Our first null hypothesis is given by Proposition 3. The corresponding alternative hypothesis is divided into two parts: First, the sender should lie less than predicted if preferences for truth-telling matter. If this is true, the best response function of the receiver in the logit-AQRE model dictates to increase

the probability with which action D is taken after message A and with which action U is taken after message B, or, to say it differently, the receiver should take more often the action *trust*.[5]

HYPOTHESIS 1: *In the Benchmark Game, the senders lie in less than fifty percent of all observations and the receivers trust the senders in more than fifty percent of all observations.*

Note that if Hypothesis 1 turns out to be correct and the senders tell the truth more often than predicted by the unique logit-AQRE, we are still left to guess the factor causing the excessive truth-telling, because it could in principle be caused by some kind of bounded rationality that is not captured by the logit-AQRE and not by preferences for truth-telling. For this reason, we introduce punishments into the original game in order to trigger subject's concerns for procedural justice. We then study how individuals react to deceptions and use these responses to indicate that preferences for truth-telling play a central role in strategic information transmission.[6]

The Punishment Game

The Punishment Game extends the Benchmark Game. Let H be the set of all histories of the Benchmark Game. Given a particular history $h = h(a, m, \theta) \in H$, the receiver reduces the payoffs of both participants to zero with probability $r(h) \in [0, 1]$ and he accepts the payoff distribution induced by the Benchmark Game with probability $1 - r(h)$. Then, both players receive their payoff. ■

Given the strategy (p_A, p_B) of the sender in the Punishment Game, $l_p(p_A, p_B)$ and $t_p(p_A, p_B)$ denote the probabilities that the sender lies and that the receiver trusts the sender's message, respectively. It is easy to calculate the set of sequential equilibria of the Punishment Game, because from a purely materialistic point of view it is never optimal for the receiver to reduce payoffs. This is different for the logit-AQRE since it allows for noisy behavior: Given $\lambda \in [0, \infty)$

[5] We equate the receivers trust(distrust) with receivers believing the payoff table to be the one (opposite) of the sender's message and playing a best response to this belief. As indicated by an anonymous referee, one way to find out whether this method is correct is to ask the receivers which payoff table they think they are facing following the sender's message and prior to their choice. We implemented this question in our robustness analysis. The resulting overall error rate was as low as 7.8%.

[6] This approach is similar to the one Fehr and Gächter (2000) who address the importance of reciprocity in public good games. Without punishments, reciprocal individuals have no possibility to enforce a positive contribution level of the selfish people, and therefore, the best they can do is not to contribute either. But if costly punishments are available, then reciprocal types have a device to enforce high contribution levels of all subjects. One can also draw obvious connections to the Dictatorship and the Ultimatum Game.

and a history $h \in H$ that insures the receiver a payoff of $y \in \{1, x\}$, the receiver reduces payoffs to zero with probability $\tilde{r}(h) = \frac{1}{1+e^{\lambda y}}$. It follows then from a backward induction argument that the probabilities $\tilde{p}_A^*$, $\tilde{p}_B^*$, $\tilde{q}_A^*$ and $\tilde{q}_B^*$ for the Punishment Game must be the same as before. Hence, the equilibrium prediction of the logit-AQRE is such that the sender (receiver) lies (trusts) again with probability one-half.

Proposition 4 *Given $\lambda \in [0, \infty)$, the logit-AQRE of the Punishment Game is such that (a) for all $h \in H$ that give the receiver a payoff of $y \in \{1, x\}$, $\tilde{r}(h)^* = \frac{1}{1+e^{\lambda y}}$ and (b) $l_p(\tilde{p}_A^*, \tilde{p}_B^*) = t_p(\tilde{p}_A^*, \tilde{p}_B^*) = \frac{1}{2}$.*

Proof: See Appendix C. □

To derive our hypothesis with respect to the punishment behavior note that the set H can be summarized by the histories h_1 = (truth,trust), h_2 = (truth,distrust), h_3 = (lie,trust), and h_4 = (lie,distrust).[7] According to Proposition 4, if receivers play the logit-AQRE, then the punishment rate after any history depends only on the payoff the receiver foregoes. On the other hand, the inequity aversion models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) take into account that some individuals care not only about their own payoff but about the whole payoff distribution. Since the payoff distribution after the histories h_2 = (truth,distrust) and h_3 = (lie,trust) is equal to $(x, 1)$, both inequity aversion and the logit-AQRE predict that the punishment rate after these two histories is the same. This constitutes our second null hypothesis.

Our alternative hypothesis is based on the findings of Brandts and Charness (2003) who have shown that receivers punish senders more often if the payoff distribution results from a deceptive message, suggesting thus that the utility attached to a particular payoff distribution also depends on how it has been reached. According to this experimental finding, individuals should punish the sender more frequently after history h_3 = (lie,trust) than after history h_2 = (truth,distrust). We do not expect any punishments after the histories h_1 = (truth,trust) and h_4 = (lie,distrust), because the receiver interpreted the message correctly and the resulting payoff distribution $(1, x)$ is favorable to him. If we reject the null hypothesis in favor of the alternative one, then a sender who lies is in more danger of being punished than a truth-teller. Thus, we hypothesize that truth-telling is enhanced in the Punishment Game with respect to the Benchmark Game and that the receivers consequently trust more in the former than in the latter.

[7] In our experimental sessions we do not ask subjects to elicit their mixed strategies, rather we derive them from the repeated observation of pure strategies. Then, since the payoff tables A and B are symmetric and the probabilities $p(A)$ and $p(B)$ are identical, we can write the set H as described. See also footnote #5.

HYPOTHESIS 2: *In the Punishment Game, the receivers punish the senders only after the histories* $h_2 = (truth, distrust)$ *and* $h_3 = (lie, trust)$ *with the punishment rate being higher after* h_3*. Moreover, the senders lie less and the receivers trust more in the Punishment than in the Benchmark Game.*

According to our main hypothesis the excessive truth-telling is caused by number of individuals that take into account social norms. To check this we perform a final consistency test on the Punishment Game. After observing the experimental results, we divide our subject pool into two different groups, one group of subjects punishing liars frequently after history h_3 = (lie,trust) and another group containing the rest of the subjects. Given this division, the third null hypothesis states that there is no difference in the level of truth-telling between the two groups when subjects play the Punishment Game as senders.[8] The corresponding alternative hypothesis, on the other hand, states that the group of subjects with a high sense of procedural justice account for most of the excessive truth-telling in the Punishment Game; that is, these subjects tell the truth very often whereas the rest of the subjects lie in about fifty percent of the occasions.

HYPOTHESIS 3: *In the Punishment Game, the group of subjects punishing liars frequently after history* $h_3 = (lie,trust)$ *accounts for most of the excessive truth-telling.*

3 Experimental Design and Procedures

We conducted our first experimental series (x=2) at the University of Edinburgh in May 2004. Since all economics students at this university have an E-mail account associated to their matriculation number, we promoted the experiment mainly via electronic newsletters. Students from other academic disciplines were recruited through flyers distributed on the campus and further announcements made on information boards. As a result, 132 undergraduate students from nearly all faculties participated in one of our experimental sessions. We organized a total of ten sessions, five on the Benchmark and five on the Punishment Game. Twelve subjects participated in the first four sessions and eighteen subjects in the fifth and last session of each treatment. No subject took part in more than one session.

To perform the experiment we employed the computer software Z-Tree developed by Fischbacher (1999). At the beginning of a session, subjects met in a

[8] We use a role rotation mechanism in our experimental sessions so that every subject plays the Punishment Game half of the time in each role. For more on this see Section 3.

computer room and took place in front of one of the computers. The computers were placed in such a way that all subjects could only look at their own screen. Next to each computer we placed a closed envelope containing instructions, a questionnaire, and a payment receipt. After subjects had filled out the questionnaire we read the instructions aloud (see Appendix D for the instructions corresponding to the Punishment Game).

Before the first round of a session, the computer randomly divided subjects into groups of six without revealing the actual matching. Thus, the students, who were all anonymous to each other, did not know who else was in their group. We informed every subject that s/he would only play against subjects belonging to the same group. Therefore, the fact that the number of subjects differed across sessions should not matter. So we implicitly divided our subject pool into a total of twenty-two groups of six subjects, eleven groups playing each treatment. In each of the fifty rounds of an experimental session the computer matched the subjects belonging to the same group into three new pairs and assigned different roles (sender or receiver) within pairs. The matchings were balanced so that after fifty rounds every subject played the game exactly ten times against each of her/his five opponents. Moreover, every subject met every opponent five times in each role. The order of the matching within a group was unknown to the subjects.

In every round, after pairs had been formed and roles had been assigned, the sender was informed of whether table A or B had been selected. Then, the sender transmitted a message from the message space $M = \{A, B\}$ telling the receiver which table corresponds to the actual payoff scheme. Afterwards, the receiver chose an action from the action set $\mathcal{A} = \{U, D\}$. This constituted the end of the round in the Benchmark Game. In the sessions corresponding to the Punishment Game, the receiver was further informed about the induced payoffs of her/his action. Finally, s/he had to decide between accepting these payoffs or reducing the payoff of both participants to zero.

At the end of a session, we called subjects one by one to step forward to the control desk for payment. In addition to the five pounds show up fee, subjects received ten pence per point in the payoff table; that is, subjects earned either 10 or 20 pence per round. As a result, the average payment in the one hour session corresponding to the Benchmark and the Punishment Game was equal to 12.5 pounds and 11.74 pounds, respectively.

The robustness analysis (x=9) was performed at Maastricht University in December 2005. Students from the economics and business faculty were able to register online for this experimental series. In total, 120 students participated in one of the ten sessions (five with respect to both treatments). Since twelve

students took part in every session, we obtained ten independent observation for both treatments. The experimental design was nearly identical to the original one, the only difference was that we asked the receivers additionally about their beliefs with respect to the actual payoff table after observing the sender's message and before taking an action. This helped us to elicit the receiver's action space (see footnote #5). Due to financial constraints we paid less money as in our first experimental series. Actually, students received for the one hour experiment twice their average points per round. As a result, the average payment was 10 Euros for students participating in the Benchmark Game and about 9.27 Euros for the ones taking part in a session corresponding to the Punishment Game.

4 Experimental Results (x=2)

4.1 Excessive Truth-Telling in the Benchmark Game

According to our first null hypothesis, the senders lie in the Benchmark Game with probability one-half. In the histogram in the left part of Figure 2, we represent the frequencies of truthful messages in the sessions corresponding to the Benchmark Game. Since a subject was exactly 25 times in the role of the sender, in equilibrium s/he should tell the truth 12.5 times. The data seems to be slightly shifted to the right of the theoretical mean, but it is not clear-cut enough to reject the null hypothesis straight away. In the right panel of Figure 2, it is possible to observe that the percentage of subjects telling the truth is extraordinarily high in the first rounds and declines over time in such a way that it stays on average just above the 50%-line predicted by the null hypothesis. We eliminate this learning effect by excluding the data from the first ten rounds in our statistical analysis.

Put Figure 2 about here [caption: Senders' Behavior ($x = 2$, Benchmark)]. The corresponding graphic files are called hist(2SB).eps (left panel) and time(2SB).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

Since subjects belonging to the same group play the Benchmark Game more than once against the other group members, actions within a given group are likely to be correlated over time. One way of obtaining independent observations is to calculate for every group the percentage of truthful messages over the last forty rounds. This procedure allows us to derive a total of eleven independent observations, one for each group. The overall percentage of truth-telling in the

last forty rounds is equal to 55.07%, a percentage significantly greater than 50% (p-value of the one-tailed Wilcoxon rank-sum test = 0.0615; p-value of the one-tailed t-test = 0.0459).

Next, we provide evidence in favor of the second part of Hypothesis 1, namely that the receivers adjust their beliefs in the correct direction and trust the senders in more than fifty percent of all occasions. To this end, we interpret the action of the receiver as the result of a maximization process involving subjective beliefs about the truthfulness of the message. For example, if a subject observes message A and takes action D afterwards, then this action reveals in our understanding that the subject trusted the sender's message. In the histogram in the left panel of Figure 3, we can clearly see that a lot of receivers trust more often than the theoretical prediction of 12.5 times. If we analyze the evolution of this percentage over time (the right panel of Figure 3), one observes that it is particularly low in the first rounds before it stabilizes well above the 50%-line.

Put Figure 3 about here [caption: Receivers' Behavior ($x = 2$, Benchmark)]. The corresponding graphic files are called hist(2RB).eps (left panel) and time(2RB).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

In the last forty rounds of the experiment the receivers trusted the senders' messages in 58.7% of all observations. This value is significantly greater than the theoretical prediction (p-value of the one-tailed Wilcoxon rank-sum test = 0.0019; p-value of the one-tailed t-test = 0.0006). Hence, we reject the null hypothesis in favor of Hypothesis 1.

In addition to their statistical significance, these deviations from the equilibrium prediction are economically relevant. Given receivers' behavior, senders' expected payoff when they send a truthful message is a 11% lower compared to the case when they lie. Subjects are thus foregoing a significant amount by telling the truth. On the other hand, excessive truth-telling also means that receivers who trust the sender's message get in expectation a 6.5% higher payoff than if they distrust.

4.2 Procedural Justice in the Punishment Game

So far we have shown the existence of excessive truth-telling in the Benchmark Game. The analysis of the Punishment Game will help us to shed light on the origin of this result. In Table 2 below we present the punishment behavior of the receivers. For consistency reasons we only consider punishments in the last

forty rounds, and therefore, we have a total of 1320 observations (11 groups of 40 rounds and 3 observations per round). The senders told the truth 740 times and lied in 580 occasions. The receivers trusted the message 520 times when the sender had told the truth before and 396 times when the sender had lied.

Put Table 2 about here [caption: Punishment Behavior ($x = 2$)].

The punishment rate is the highest, more than 25%, after history h_3 = (lie,trust). We use the normal approximation of the binomial distribution in order to establish that this proportion is significantly greater than zero (p-value of the one-tailed Z-test < 0.0001). We also find, as expected, that the punishment rate after history h_2 = (truth,distrust) is significantly greater than zero. We attribute the positive punishment rate after history h_4 = (lie,distrust) to mistakes made by some subjects. Yet, our main prediction is confirmed: The willingness to punish the sender depends on whether or not a payoff distribution has been reached by means of a deceptive message, because the punishment rate after history h_3 = (lie,trust) is greater than the one after history h_2 = (truth,distrust). A test of equal proportions confirms this observation (p-value of the one-tailed Z-test < 0.0001).

We investigate next whether subjects behave consistently across the two treatments. The histogram in left panel of Figure 4 looks quite similar to the one corresponding to sender's behavior in the Benchmark Game although it seems that the shift to the right from the theoretical mean has increased. In the right panel of Figure 4, we observe that the percentage of subjects telling the truth is quite high in the first rounds and declines over time, a behavior we have already encountered before. Nevertheless, in the latter rounds there are now less values below the fifty percent line.

Put Figure 4 about here [caption: Senders' Behavior ($x = 2$, Punishment)]. The corresponding graphic files are called hist(2SP).eps (left panel) and time(2SP).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

The percentage of subjects telling the truth in the last forty rounds of the Punishment Game is equal to 56.29%. This percentage is significantly greater than the equilibrium prediction (p-value of the one-tailed Wilcoxon rank-sum test = 0.0499; p-value of the one-tailed t-test = 0.0343), but it is not significantly greater than the corresponding value for the Benchmark Game (p-value of the one-tailed Wilcoxon rank-sum test = 0.409; p-value of the one-tailed t-test = 0.3855).

The picture looks quite different if we compare the receivers' behavior across the two treatments. The histogram in the left panel of Figure 5 indicates that the receivers trust more in the Punishment than in the Benchmark Game. This intuition is confirmed in the right panel of Figure 5, because the percentage of receivers trusting the sender seems to increase over time and stays well above the equilibrium prediction. On the aggregate, the percentage of trustful receivers in the last forty rounds is equal to 69.3%. This percentage is significantly greater than the corresponding value of the Benchmark Game (p-value of the one-tailed Wilcoxon rank-sum test = 0.0031; p-value of the one-tailed t-test = 0.0036).

Put Figure 5 about here [caption: Receivers' Behavior ($x = 2$, Punishment)]. The corresponding graphic files are called hist(2RP).eps (left panel) and time(2RP).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

Again, the economic significance of the observed departures from the equilibrium predictions is considerable. Not only are punishments relatively frequent, but also lying becomes the optimal strategy given receivers' trust-worthiness and punishments rates: By telling the truth, senders receive in expectation a 4.8% lower payoff than if they lie; whereas receivers gain in expected terms, and without taking into account the punishments, a 8.7% by trusting the sender's message.

To summarize: We have confirmed the importance of procedural justice in this sender-receiver game. People frequently punish when they are deceived. Moreover, the introduction of punishments into the Benchmark Game seems to induce the receivers to believe that the senders will often tell the truth in order to avoid a possible moral outrage caused by deceptive messages. But on the contrary, when subjects play as senders they seem to consider the punishment as an incredible threat, because they barely change their behavior with respect to the original set-up.

4.3 Morally Consistent Behavior

So far we have laid down the ground for our main result, namely that the tension between incentives and normative social behavior is the driving force behind the excessive truth-telling. After observing the experimental results, we divide our subject pool into two groups, one group containing all those subjects who punish liars frequently after history h_3 = (lie,trust) and another group containing the rest of subjects. We obtain this division in the following way: In the last forty rounds of an experimental session corresponding to the Punishment Game every

subject is twenty times in the role of the receiver. Since the sender lies with probability 0.437 and the receiver trusts the message with probability 0.694, every subject plays, in expected terms, the history h_3 = (lie,trust) 6.06 times in the role of the receiver. The punishment rate after h_3 is equal to 0.2525, and therefore, every subject is expected to punish the sender 1.53 times. Hence, all subjects that punish the sender in at least three occasions after h_3 reveal serious concerns for procedural justice. This condition is met by fifteen out of sixty-six subjects. Not surprisingly, this group of subjects accounts for 90% of all punishments after h_3. Moreover, from the 110 times one subject belonging to this group played history h_3 in the role of the receiver, the sender was punished in 90 occasions (this is equal to a punishment rate of 81.81%) whereas the rest of the individuals punished the sender only 3.16% of all observations after this history.

Given this classification, the role rotation mechanism allows us to study how these fifteen subjects behave in the Punishment Game. On the aggregate, they tell the truth in 70.66% of all observations. This probability is significantly greater than 56.29%, the percentage of truth-telling corresponding to the whole subject pool (p-value of the one-tailed Z-test < 0.0001). The rest of the subjects, on the other hand, tell the truth in only 52.05% of the cases, a percentage not significantly greater than the equilibrium prediction (p-value of the one-tailed Z-test $= 0.0945$). Therefore, we reject the third null hypothesis - the percentage is the same for both groups of subjects - in favor of Hypothesis 3.[9] This result indicates the existence of what we refer to as *morally consistent behavior:* Individuals with a strong notion of procedural justice behave consistently across roles and are responsible for nearly all the information transmitted by the senders. This interpretation is further strengthened if we analyze how the beliefs of these two groups vary. Subjects with a serious notion for procedural justice trust the senders' message in 86% of all occasions, whereas the rest of subjects do so only in 64.51%.

5 Robustness Analysis (x=9)

How do our results change as we increase the inequality in the payoff distribution between sender and receiver while maintaining incentives the same? To address this question, we run an additional experimental series considering the case when $x = 9$. According to our fourth null hypothesis this change should not

[9] A different consistency test puts into one group all individuals that never punish a sender. It turns out that this group tells the truth with probability 0.484, a value that is not significantly different from logit-AQRE. The group consisting of all subject that punish a sender at least once, on the other hand, tells the truth with probability 0.684.

affect the experimental results. Next, we are going to develop our alternative hypothesis, which points in a different direction.

Suppose that participants play the Benchmark Game. If $x = 2$ and the sender thinks that the receiver will take the action trust (this action could be defined as the social standard), then her optimal strategy would be to lie and the associated cost of telling the truth is 1. This cost is rather small in comparison to the case when $x = 9$ and suggests that senders lie more when the *price* of truth-telling increases (see e.g. Zwick and Chen (1999)). According to the best response function of the logit-AQRE, the receivers should then also trust less. What we expect to see is that an increase in inequality leads to a clear shift towards the logit-AQRE.

Perhaps surprisingly, we conjecture the opposite result in the Punishment Game. Since the payoff distributions are now less "fair", models of inequity aversion predict that the receivers punish more whenever they get the low payoff (this corresponds to the history h_2 and h_3). At the same time, the receivers may feel even more deceived after history h_3 according to the notion of procedural justice, because they see that the sender gets away with an even larger share of the total payoff thanks to her unethical behavior. This should lead to additional punishments, and therefore, the difference in the punishment rates between h_2 and h_3 should increase in comparison to our first experimental series. As a consequence, the level of truth-telling and trust should also increase.

HYPOTHESIS 4: *In the Benchmark Game, senders lie less and receivers trust more when $x = 2$. In the Punishment Game, the difference in the punishment rate between $h_3 =$ (lie,trust) and $h_2 =$ (truth,distrust) is higher when $x = 9$. As a consequence, senders tell the truth more often and receivers trust more when $x = 9$.*

5.1 Equilibrium Play in the Benchmark Game

Hypothesis 4 states that there should be a drift towards the logit-AQRE in the Benchmark Game. Figure 6 provides a first indication that this is true for the sender. The data in the histogram in the left panel of Figure 6 seems to be centered around the equilibrium prediction. Additionally, we see in the right panel of Figure 6 that the level of truth-telling converges quickly towards the equilibrium prediction and oscillates around it afterwards. These observations are confirmed by the statistics: the percentage of truth-telling over the last forty rounds is equal to 50.6%. This value is significantly smaller than the corresponding percentage for the first experimental series of 55.07% (p-value of the one-tailed Wilcoxon rank-sum test = 0.1105; p-value of the one-tailed t-test

= 0.111) and not significantly greater than the equilibrium prediction (p-value of the one-tailed Wilcoxon rank-sum test = 0.4549; p-value of the one-tailed t-test = 0.3803).

Put Figure 6 about here [caption: Senders' Behavior ($x = 9$, Benchmark)]. The corresponding graphic files are called hist(9SB).eps (left panel) and time(9SB).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

Next, we analyze whether receivers also behaved according to the equilibrium prediction; is the percentage of *trust* lower than before? The histogram in left panel of Figure 7 and the evolution of the percentage of trust in the right panel of the same Figure do not seem to support this conjecture. Both look very much the same as before. The percentage of trust over the last forty rounds is now equal to 58.8%. This value is significantly greater than the equilibrium prediction (p-value of the one-tailed Wilcoxon rank-sum test = 0.0029; p-value of the one-tailed t-test = 0.001). Recall that in the first series it was 58.7%. These two values are obviously not significantly different from each other (p-value of the one-tailed Wilcoxon rank-sum test = 0.5808; p-value of the one-tailed t-test = 0.5009).

Put Figure 7 about here [caption: Receivers' Behavior ($x = 9$, Benchmark)]. The corresponding graphic files are called hist(9RB).eps (left panel) and time(9RB).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

Our analysis reveals that the excessive truth-telling vanishes as more inequality is introduced into the payoff distribution. This is consistent with the results found in Gneezy (2005) because senders are now able to achieve a higher gain from deceiving the receiver, or alternatively, according to Zwick and Chen (1999), because the relative price of truth-telling has increased and its demand has consequently dropped. We have predicted this result in Hypothesis 4. However, it is surprising that, although they should move together, excessive trustworthiness does not vanish. We have no clear explanation for this. On the one hand, senders are basically randomizing and this leaves receivers indifferent between trusting or not the sender's message (the receiver's expected payoff from trusting the message is only a 0.8% higher than from not trusting). Hence, receivers might be just conforming to a social standard or taking the simplest strategy they can think of. On the other hand, it may well be the case that

receivers held wrong beliefs about senders' behavior and expected higher levels of truth-telling.[10]

5.2 Enhanced Procedural Justice in the Punishment Game

Even though the excessive truth-telling has apparently vanished, it is still worthwhile to analyze the corresponding Punishment Game. First, because it constitutes an obvious robustness check of our former results. But more importantly because it can put at test our claim that procedural justice becomes more focal as the inequality of the payoff distribution increases. In Table 3 we present the punishment behavior of the receivers. We have a total of 1200 observations (10 groups of 40 rounds and 3 observations per round). The senders told the truth 675 times and lied in 525 occasions. The receivers trusted the message 490 times when the sender had told the truth and 194 times when the sender had lied.

Put Table 3 about here [caption: Punishment Behavior ($x = 9$)].

Our main observation is that the punishment rate increased from 25.2% to 42.8% after history h_3 = (lie,trust) and from 5.4% to 13.4% after history h_2 = (truth,distrust). According to the logit-AQRE none of the two values should have changed. Inequity aversion, on the other hand, explains why the receivers punish the sender more often after these histories, but it cannot account for the fact that difference in the punishment rates between history h_3 and h_2 has increased from 19.8% to over 29.4%. Since the latter value is significantly greater than the corresponding value in our first experimental series (p-value of the one-tailed Z-test < 0.0001), we reject the null hypothesis that the value of x does not influence the difference in the punishment rate in favor of Hypothesis 4 that procedural justice has become even more important. Furthermore, observe that there are no punishments any more after history h_4 = (lie,distrust).

Now, we want to study whether this result has any effect on the aggregate level of truth-telling and the aggregate level of trust in the Punishment Game. In the histogram in the left panel of Figure 8 we observe that the senders behave very heterogeneously, whereas the right panel reveals that the percentage of truth-telling per round stays rather constantly above the equilibrium prediction. The percentage of truth-telling over the last forty rounds is equal to 56.4% whereas the corresponding value in our first experimental series was 56.3%. The former value is not significantly greater than the latter (p-value of the one-tailed

[10] Because levels of truth-worthiness remain the same compared to the case of $x = 2$, the expected loss from telling the truth instead of sending a deceptive message is again 11%.

Wilcoxon rank-sum test = 0.5391; p-value of the one-tailed t-test = 0.4876), but obviously it is significantly greater than the equilibrium prediction (p-value of the one-tailed Wilcoxon rank-sum test = 0.0186; p-value of the one-tailed t-test = 0.0167). Thus, we accept the null hypothesis that the value of x does not effect the senders' behavior in the Punishment Game.

Put Figure 8 about here [caption: Senders' Behavior ($x = 9$, Punishment)]. The corresponding graphic files are called hist(9SP).eps (left panel) and time(9SP).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

However, the reader should note that the introduction of punishments in the second experimental series does enhance truth-telling compared to the Benchmark Game. The percentage of truthful messages remains above the equilibrium prediction despite the fact that increased inequality made the excessive truth-telling vanish in the Benchmark Game. Two countervailing forces seem to be at work here: On the one hand, the higher punishment rates when x=9 certainly induce senders to tell the truth more often but at the same time the gain from lying (or alternatively, the price of truth-telling) has increased. Even though the net effect is positive in comparison with the Benchmark Game when x=9, the overall effect nullifies in comparison with the level of truth-telling in the Punishment Game when x=2.

Now, we analyze whether the behavior of the receiver remains robust to an increase in inequality. Figure 9 bears a lot of similarities with Figure 5, where we presented our results with respect to the receivers' behavior in the original Punishment Game. Now, the percentage of trust over the last forty rounds is equal to 71.1%. This value is significantly greater than the equilibrium prediction of 50% (p-value of the one-tailed Wilcoxon rank-sum test = 0.001; p-value of the one-tailed t-test < 0.0001), but not significantly greater than 69.3%, the percentage corresponding to the original set-up (p-value of the one-tailed Wilcoxon rank-sum test = 0.2461; p-value of the one-tailed t-test = 0.344). Thus, we accept the null hypothesis that the value of x does not influence the receiver's behavior in the Punishment Game.

Put Figure 9 about here [caption: Receivers' Behavior ($x = 9$, Punishment)]. The corresponding graphic files are called hist(9RP).eps (left panel) and time(9RP).eps (right panel). They should be scaled down to 47.5% and placed next to each other.

An interesting point to make is that the wider gap between the punishment rates for histories h_3 = (lie,trust) and h_2 = (truth,distrust) now makes telling

the truth the optimal strategy for senders. A truthful message increases their expected payoff by a 9.1 % compared to lying; meanwhile, trusting the sender (no punishments considered) enhances receivers' expected payoff by a 8.9%.

5.3 Morally Consistent Behavior

Our last aim is to study whether the excessive truth-telling in the Punishment Game can again be explained in terms of preferences for truth-telling. Since the sender lies with probability 0.436 and the receiver trusts the message with probability 0.711, every subject plays, in expected terms, the history h_3 = (lie,trust) 6.2 times in the role of the receiver. The punishment rate after h_3 is equal to 0.4286, and therefore, every subject is expected to punish the sender 2.65 times. Hence, all subjects that punish the sender in at least five occasions after h_3 reveal serious concerns for procedural justice. This condition is met by seventeen out of sixty subjects. This group of subjects accounts for 78.8% of all punishments after h_3 and punishes the sender with probability 0.8282 after this history. The rest of the individuals punish the sender only in 15.34% of all observations after this history. This value, by the way, is more or less the same percentage as the overall punishment rate after history h_2.

On the aggregate, the group of individuals with serious concerns for procedural justice tell the truth in 69.4% of all observations. This probability is significantly greater than 56.4%, the percentage of truth-telling corresponding to the whole subject pool (p-value of the one-tailed Z-test < 0.0001). Since the rest of the subjects tell the truth in only 51.20% of the cases, we can conclude that our main hypothesis is robust to the change in the potential payoff distribution. Further indications for this interpretation stem from the fact that the group of "moral" individuals trusts the sender in 84.7% of all cases whereas the rest of the subjects do so only with probability 0.656. Finally, the group of individuals who never punish tells the truth with probability 0.485, whereas the group of subjects that punish in the last forty rounds at least once tells the truth with probability 0.6287.

6 Conclusion

Communication is the most natural way how to exchange information. Experimental studies such as the ones of Duffy and Feltovich (2002) and (2006) have shown that individuals are able to achieve Pareto improving allocations by means of cheap talk. In particular, the authors show that if subjects announce to cooperate in the Prisoner's Dilemma, then this message often reflects

the truth. Moreover, receivers reciprocate and cooperate as well so the Pareto-efficient outcome is sometimes implemented. On the other hand, Crawford (2003) shows that in some sequential equilibria of a sender-receiver game a rational individual can feint a boundedly rational one. These results raise some questions. In which situations can the receiver trust the senders' messages? And why do the senders transmit truthful messages if incentives suggest otherwise? Our aim was to show that the overcommunication phenomenon of Cai and Wang (2005) is not necessarily due to a lack of sophistication or rationality but results from the fact that some individuals take into account social norms such as truth-telling.

To this end, we studied the behavior of a group of subjects in a simple game of strategic information transmission. We showed in the first step of our analysis that in the Benchmark Game, senders tell the truth more often than predicted by the unique logit-AQRE. Then, we introduced punishments and established that, in accordance with the results of Brandts and Charness (2003), the willingness to punish the sender is higher after a deceptive message. Finally, we sustained our main hypothesis by showing that if we subtract from our subject pool the group of subjects who punish liars frequently after a deceptive message, then that very same group tells the truth very often whereas the rest of the subjects behave roughly according to the standard equilibrium prediction. Thus, if moral subjects are excluded, the excessive truth-telling vanishes. In our robustness analysis we increased the inequality of the payoff distribution leaving incentives the same. Our two main findings are that there is no excessive truth-telling any more in the Benchmark Game and that the notion of procedural justice becomes even more important in the Punishment Game.

In further research we intend to explore the existence and extent of the *morally consistent behavior* that we have uncovered in the present paper. The existence of moral individuals who reject material incentives to misbehave opens some fascinating questions: What are the implications on mechanism design? Or on the organization of the firm? And on the elaboration of policy prescriptions?

References

[1] Alingham, M., Sandmo, A., 1972. Income tax evasion: a theoretical analysis. J. Public Econ. 1, 323-338.

[2] Blume, A., DeJong, D., Kim, Y., Sprinkle, G., 1998. Experimental evidence on the evolution of meaning of messages in sender-receiver games. Amer. Econ. Rev. 88, 1323-1339.

[3] Bolton, G., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. Amer. Econ. Rev. 90, 166-193.

[4] Brandts, J., Charness, G., 2003. Truth or consequence: an experiment. Management Sci. 49, 116-130.

[5] Cai, H., Wang, J., 2006. Overcommunication in strategic information transmission games. Games Econ. Behav. 95, 384-394.

[6] Costa-Gomes, M., Crawford, V., Broseta, B., 2001. Cognition and behavior in normal form games: an experimental study. Econometrica 69, 1193-1235.

[7] Crawford, V., 2003. Lying for strategic advantages: rational and boundedly rational misrepresentations of intentions. Amer. Econ. Rev. 93, 133-149.

[8] Crawford, V., Sobel, J., 1982. Strategic information transmission. Econometrica 50, 1431-1451.

[9] Dickhaut, J., McCabe, K., Mukherji, A., 1995. An experimental study of strategic information transmission. Econ. Theory 6, 389-403.

[10] Duffy, J., Feltovich, N., 2002. Do actions speak louder than words? Observation vs. cheap talk as coordination devices. Games Econ. Behav. 39, 1-27.

[11] Duffy, J., Feltovich, N., 2006. Words, deeds and lies: strategic behavior in games with multiple signals. Rev. Econ. Stud. 73, 669-688.

[12] Fehr, E., Schmidt, K., 1999. A theory of fairness competition, and cooperation. Quat. J. Econ. 114, 817-864.

[13] Fischbacher, U., 1999. Z-Tree - Zurich toolbox for readymade economic experiments - experimenter's manual. Working Paper Nr. 21, Institute of Empirical Research in Economics, Zurich University.

[14] Galor, E., 1985. Information sharing in oligopoly. Econometrica 53, 329-343.

[15] Gneezy, U., 2005. Deception: the role of consequences. Amer. Econ. Rev. 95, 384-394.

[16] Heidhues, P., Lagerlof, J., 2003. Hiding information in electoral competition, Games Econ. Behav. 42, 48-74.

[17] McKelvey, R., Palfrey, T., 1995. Quantal response equilibria in normal form games, Games Econ. Behav. 10, 6-38.

[18] McKelvey, R., Palfrey, T., 1998. Quantal response equilibria in extensive form games, Exper. Econ. 1, 9-41.

[19] Morgan, J., Stocken, P., 2003. An analysis of stock recommendations. RAND J. Econ. 34, 183-203.

[20] Nagel, R., 1995. Unraveling in guessing games: an experimental study. Amer. Econ. Rev. 85, 1313-1326.

[21] Sen, A., Maximization and the act of choice. Econometrica 65, 745-779.

[22] Zwick, R., Chen, X., What price fairness? A bargaining study. Management Sci. 44, 119-141.

Appendix A: Proof of Proposition 1

Recall that p_A (or p_B, respectively) denotes the probability that the sender submits message A when the actual payoff scheme is represented by table A (or table B, respectively). We divide our analysis into three different cases.

Case 1: Suppose that $0 < p_A + p_B < 2$. We derive the best response correspondence for the receiver who takes the strategy (p_A, p_B) of the sender as given. Suppose that the sender transmits message A. By sequential rationality the receiver updates his beliefs according to Bayes' rule, and therefore, he thinks that the probability $\mu(A|A) \equiv \mu_A$ (e.g. the true payoff scheme is given by table A conditional on message A) is equal to

$$\mu_A = \frac{p(m=A|\theta=A)p(A)}{p(m=A)} = \frac{0.5p_A}{0.5p_A+0.5p_B} = \frac{p_A}{p_A+p_B}.$$

Given μ_A, the receiver chooses q_A (the probability to take action U conditional on message A) in order to

$$\max_{q_A} \quad (\mu_A (q_A + x(1-q_A)) + (1-\mu_A)(xq_A + 1 - q_A)).$$

This maximization problem is equivalent to

$$\max_{q_A} \quad (1 + (x-1)\mu_A + (x-1)q_A(1-2\mu_A)),$$

and therefore, the best response correspondence for the receiver is

$$q_A^*(\mu_A) = \begin{cases} 1 & \text{if } \mu_A < \frac{1}{2} \\ [0,1] & \text{if } \mu_A = \frac{1}{2} \\ 0 & \text{if } \mu_A > \frac{1}{2}, \end{cases} \quad \text{or} \quad q_A^*(p_A, p_B) = \begin{cases} 1 & \text{if } p_A < p_B \\ [0,1] & \text{if } p_A = p_B \\ 0 & \text{if } p_A > p_B. \end{cases}$$

If, on the other hand, the sender submits message B, then the belief that the actual payoff scheme is represented by table A, μ_B, is equal to

$$\mu_B = \frac{p(m{=}B|\theta{=}A)p(B)}{p(m{=}B)} = \frac{0.5(1-p_A)}{0.5(1-p_A)+0.5(1-p_B)} = \frac{1-p_A}{2-p_A-p_B}.$$

Given μ_B, the receiver chooses q_B (the probability to take action U conditional on message B) in order to

$$\max_{q_B} \quad (\mu_B (q_B + x (1 - q_B)) + (1 - \mu_B) (xq_B + 1 - q_B)) .$$

This maximization problem is equivalent to

$$\max_{q_B} \quad (1 + (x - 1)\mu_B + (x - 1)q_B(1 - 2\mu_B)) ,$$

and therefore, the best response correspondence of the receiver is

$$q_B^* (\mu_B) = \begin{cases} 1 & \text{if } \mu_B < \frac{1}{2} \\ [0,1] & \text{if } \mu_B = \frac{1}{2} \\ 0 & \text{if } \mu_B > \frac{1}{2}, \end{cases} \qquad \text{or} \qquad q_B^* (p_A, p_B) = \begin{cases} 1 & \text{if } p_A > p_B \\ [0,1] & \text{if } p_A = p_B \\ 0 & \text{if } p_A < p_B. \end{cases}$$

Next, we calculate the optimal mixed strategy (p_A^*, p_B^*) for the sender. To do so we consider three different cases:

Case A: Suppose that $p_A^* < p_B^*$. Then, it follows from the optimal behavior of the receiver that $q_A^*(p_A^*, p_B^*) = 1$ and $q_B^*(p_A^*, p_B^*) = 0$. Thus, the optimal strategy (p_A^*, p_B^*) must be the solution of the following maximization problem: Choose p_A and p_B in order to

$$\max_{p_A, p_B} \quad 0.5 (xp_A + p_B + 1 - p_A + x (1 - p_B)) .$$

This maximization problem is equivalent to

$$\max_{p_A, p_B} \quad 0.5(x - 1) (1 + p_A - p_B) .$$

But the solution to this problem is such that $p_A^* = 1$ and $p_B^* = 0$, and therefore, we have reached a contradiction. We conclude that there does not exist any equilibrium in which $p_A^* < p_B^*$.

Case B: Suppose that $p_A^* > p_B^*$. Then, it follows from the optimal behavior of the receiver that $q_A^*(p_A^*, p_B^*) = 0$ and $q_B^*(p_A^*, p_B^*) = 1$, and therefore, the optimal strategy (p_A^*, p_B^*) must be the solution of the following maximization problem: Choose p_A and p_B in order to

$$\max_{p_A, p_B} \quad 0.5 (p_A + x (1 - p_A) + xp_B + 1 - p_B) .$$

This maximization problem is equivalent to

$$\max_{p_A,p_B} \quad 0.5(x-1)\left(1-p_A+p_B\right).$$

But the solution to this problem is such that $p_A^*=0$ and $p_B^*=1$, and therefore, we have reached a contradiction. We conclude that there does not exist any equilibrium in which $p_A^*>p_B^*$.

Case C: Suppose that $p_A^*=p_B^*$. Then, it follows from the best response correspondences of the receiver that $q_A^*\in[0,1]$ and $q_B^*\in[0,1]$. Thus, the sender faces the problem

$$\begin{aligned}\max_{p_A,p_B} \quad & 0.5p_A\left(xq_A+1-q_A\right)+0.5\left(1-p_A\right)\left(xq_B+1-q_B\right)+\\ & 0.5p_B\left(q_A+x\left(1-q_A\right)\right)+0.5\left(1-p_B\right)\left(q_B+x\left(1-q_B\right)\right),\end{aligned}$$

a problem that is equivalent to

$$\max_{p_A,p_B} \quad 0.5(x-1)\left(1+p_A\left(q_A-q_B\right)+p_B\left(q_B-q_A\right)\right).$$

Hence, the best response correspondences for the sender are

$$p_A^*\left(q_A,q_B\right)=\begin{cases}1 & \text{if } q_A>q_B\\ [0,1] & \text{if } q_A=q_B\\ 0 & \text{if } q_A<q_B\end{cases} \quad \text{and} \quad p_B^*\left(q_A,q_B\right)=\begin{cases}1 & \text{if } q_A<q_B\\ [0,1] & \text{if } q_A=q_B\\ 0 & \text{if } q_A>q_B.\end{cases}$$

From inspection we see that the set of mixed strategies $(p_A^*,p_B^*;q_A^*,q_B^*)=(p,p;q,q)$, where $p\in(0,1)$ and $q\in[0,1]$, can be sustained as equilibrium strategies. Finally, one can easily check that the corresponding beliefs are such that $\mu_A^*=\mu_B^*=\frac{1}{2}$.

Case 2: Suppose that $p_A^*=p_B^*=0$. Observe from the best correspondence of the sender in case 1.C that $p_A^*=p_B^*=0$ can only be sustained as an equilibrium strategy if $q_A^*=q_B^*$. Moreover, $p_A^*=p_B^*=0$ implies that $\mu_B^*=\frac{1}{2}$ and $q_B^*(\mu_B^*)\in[0,1]$. Since the sequential game we study consists of two-periods and the cardinality of the action space of both players is equal to two, any belief $\mu_A\in[0,1]$ is consistent. Yet, we obtain from the best response correspondence $q_A^*(\mu_A)$ in case 1 that $q_A^*=q_B^*$ if and only if $\mu_A=\frac{1}{2}$. Therefore, we conclude that the set of mixed strategies $(p_A^*,p_B^*;q_A^*,q_B^*)=(0,0;q,q)$, where $q\in[0,1]$, together with the belief system $\mu_A^*=\mu_B^*=\frac{1}{2}$ constitutes a set of sequential equilibria.

Case 3: Suppose that $p_A^*=p_B^*=1$. Observe from the best correspondence of the sender in case 1.C that $p_A^*=p_B^*=1$ can only be sustained as an equilibrium strategy if $q_A^*=q_B^*$. Moreover, $p_A^*=p_B^*=1$ implies that $\mu_A^*=\frac{1}{2}$ and $q_A^*(\mu_A^*)\in$

$[0,1]$. Although any belief $\mu_B \in [0,1]$ is consistent, we obtain from the best response correspondence $q_B^*(\mu_B)$ in case 1 that $q_A^* = q_B^*$ if and only if $\mu_B = \frac{1}{2}$. Therefore, we conclude that the set of mixed strategies $(p_A^*, p_B^*; q_A^*, q_B^*) = (1,1;q,q)$, where $q \in [0,1]$, together with the belief system $\mu_A^* = \mu_B^* = \frac{1}{2}$ constitutes a set of sequential equilibria. □

Appendix B: Proof of Proposition 2

Remember that $\tilde{p}_A$ (or, $\tilde{p}_B$ respectively) denotes the probability in the logit-AQRE that the sender submits message A when the actual payoff scheme is represented by table A (or, table B respectively). Similarly, $\tilde{q}_A$ (or, $\tilde{q}_B$ respectively) indicates the probability that the receiver takes action U upon message A (or, message B respectively). Given $\lambda \in [0,\infty)$, we yield that

$$\tilde{p}_A = \frac{e^{\lambda u(A|A)}}{e^{\lambda u(A|A)}+e^{\lambda u(B|A)}} \quad \text{and} \quad \tilde{p}_B = \frac{e^{\lambda u(A|B)}}{e^{\lambda u(A|B)}+e^{\lambda u(B|B)}},$$

where (a) $u(A|A) = x\tilde{q}_A + (1-\tilde{q}_A) = 1 + (x-1)\tilde{q}_A$ is the expected payoff of sending message A if the payoff scheme is represented by table A, (b) $u(B|A) = x\tilde{q}_B + (1-\tilde{q}_B) = 1 + (x-1)\tilde{q}_B$ is the expected payoff of message B in state A, (c) $u(A|B) = \tilde{q}_A + x(1-\tilde{q}_A) = x - (x-1)\tilde{q}_A$ represents the expected payoff of message A when table B actually represents payoffs and (d) $u(B|B) = \tilde{q}_B + x(1-\tilde{q}_B) = x - (x-1)\tilde{q}_B$ denotes the expected payoff of sending message B in state B. It is then easy to verify that

$$\tilde{p}_A(\tilde{q}_A, \tilde{q}_B) = \frac{1}{1+e^{\lambda(x-1)(\tilde{q}_B - \tilde{q}_A)}} \quad \text{and} \quad \tilde{p}_B(\tilde{q}_A, \tilde{q}_B) = \frac{1}{1+e^{\lambda(x-1)(\tilde{q}_A - \tilde{q}_B)}}.$$

Notice that $\tilde{p}_A + \tilde{p}_B = 1$. So far, we have expressed the strategy of the sender as a function of the receivers's strategy. In the next step, we calculate the best response functions for the receiver. We obtain that

$$\tilde{q}_A = \frac{e^{\lambda v(U|A)}}{e^{\lambda v(U|A)}+e^{\lambda v(D|A)}} \quad \text{and} \quad \tilde{q}_B = \frac{e^{\lambda v(U|B)}}{e^{\lambda v(U|B)}+e^{\lambda v(D|B)}},$$

where (a) $v(U|A) = \tilde{p}_A + x\tilde{p}_B = 1 + (x-1)\tilde{p}_B$ is the expected payoff of taking action U upon message A, (b) $v(D|A) = x\tilde{p}_A + \tilde{p}_B = 1 + (x-1)\tilde{p}_A$ is the expected payoff of action D when the sender says that the actual payoff table is A, (c) $v(U|B) = (1-\tilde{p}_A) + x(1-\tilde{p}_B) = x - (x-1)\tilde{p}_B$ represents the expected payoff of action A upon message B and (d) $v(D|B) = x(1-\tilde{p}_A) + (1-\tilde{p}_B) = x - (x-1)\tilde{p}_A$ denotes the expected payoff of sending message B in state B. We can then verify that

$$\tilde{q}_A(\tilde{p}_A, \tilde{p}_B) = \frac{1}{1+e^{\lambda(x-1)(\tilde{p}_A - \tilde{p}_B)}} \quad \text{and} \quad \tilde{q}_B(\tilde{p}_A, \tilde{p}_B) = \frac{1}{1+e^{\lambda(x-1)(\tilde{p}_B - \tilde{p}_A)}}.$$

Up to now we have identified a system of four linear equations with the unknowns $\tilde{p}_A$, $\tilde{p}_B$, $\tilde{q}_A$, and $\tilde{q}_A$. Since $\tilde{p}_B = 1 - \tilde{p}_A$ and $\tilde{q}_B = 1 - \tilde{q}_A$, we can reduce the four equations to two:

$$\tilde{p}_A(\tilde{q}_A) = \frac{1}{1+e^{\lambda(x-1)(1-2\tilde{q}_A)}} \quad \text{and} \quad \tilde{q}_A(\tilde{p}_A) = \frac{1}{1+e^{\lambda(x-1)(2\tilde{p}_A-1)}}.$$

It is easy to see that $\tilde{p}_A^* = \tilde{q}_A^* = \frac{1}{2}$ is the only solution of the two equations above. It follows immediately that $\tilde{p}_B^* = \tilde{q}_B^* = \frac{1}{2}$ and $\tilde{\mu}_A^* = \tilde{\mu}_B^* = \frac{1}{2}$. □

Appendix C: Proof of Proposition 4

Fix the parameter $\lambda \in [0, \infty)$. Notice that if the history $h \in H$ is such that the receiver gets a payoff of 1, then the receiver will punish the sender with probability $\tilde{r}(h)^* = \frac{1}{1+e^{\lambda}}$. Similarly, if he gets a payoff of x, then he will punish the sender with probability $\tilde{r}(h)^* = \frac{1}{1+e^{\lambda x}}$. Knowing this, we can now apply a backward induction argument to solve for the optimal probabilities $\tilde{p}_A^*$, $\tilde{p}_B^*$, $\tilde{q}_A^*$ and $\tilde{q}_B^*$ in the very same way as before. For the sender we find that

$$\tilde{p}_A = \frac{e^{\lambda u(A|A)}}{e^{\lambda u(A|A)}+e^{\lambda u(B|A)}} \quad \text{and} \quad \tilde{p}_B = \frac{e^{\lambda u(A|B)}}{e^{\lambda u(A|B)}+e^{\lambda u(B|B)}},$$

where (a) $u(A|A) = x\tilde{q}_A\frac{e^{\lambda x}}{1+e^{\lambda x}} + \frac{e^{\lambda}}{1+e^{\lambda}}(1-\tilde{q}_A) = \frac{e^{\lambda}}{1+e^{\lambda}} + \tilde{q}_A\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)$ is the expected payoff of sending message A if the payoff scheme is represented by table A, (b) $u(B|A) = x\tilde{q}_B\frac{e^{\lambda x}}{1+e^{\lambda x}} + \frac{e^{\lambda}}{1+e^{\lambda}}(1-\tilde{q}_B) = \frac{e^{\lambda}}{1+e^{\lambda}} + \tilde{q}_B\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)$ is the expected payoff of message B in state A, (c) $u(A|B) = \tilde{q}_A\frac{e^{\lambda}}{1+e^{\lambda}} + x(1-\tilde{q}_A)\frac{e^{\lambda x}}{1+e^{\lambda x}} = x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \tilde{q}_A\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)$ represents the expected payoff of message A when table B actually represents payoffs and (d) $u(B|B) = \tilde{q}_B\frac{e^{\lambda}}{1+e^{\lambda}} + x(1-\tilde{q}_B)\frac{e^{\lambda x}}{1+e^{\lambda x}} = x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \tilde{q}_B\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)$ denotes the expected payoff of sending message B in state B. It is then easy to verify that

$$\tilde{p}_A(\tilde{q}_A, \tilde{q}_B) = \frac{1}{1+e^{\lambda\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)(\tilde{q}_B - \tilde{q}_A)}}$$

and

$$\tilde{p}_B(\tilde{q}_A, \tilde{q}_B) = \frac{1}{1+e^{\lambda\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)(\tilde{q}_A - \tilde{q}_B)}}.$$

Notice that $\tilde{p}_A + \tilde{p}_B = 1$. So far, we have expressed the strategy of the sender as a function of the receivers's strategy. In the next step, we calculate the best response functions for the receiver. We obtain that

$$\tilde{q}_A = \frac{e^{\lambda v(U|A)}}{e^{\lambda v(U|A)}+e^{\lambda v(D|A)}} \quad \text{and} \quad \tilde{q}_B = \frac{e^{\lambda v(U|B)}}{e^{\lambda v(U|B)}+e^{\lambda v(D|B)}},$$

where (a) $v(U|A) = \tilde{p}_A\frac{e^{\lambda}}{1+e^{\lambda}} + x\tilde{p}_B\frac{e^{\lambda x}}{1+e^{\lambda x}} = \frac{e^{\lambda}}{1+e^{\lambda}} + \tilde{p}_B\left(x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^{\lambda}}{1+e^{\lambda}}\right)$ is the expected payoff of taking action U upon message A, (b) $v(D|A) = x\tilde{p}_A\frac{e^{\lambda x}}{1+e^{\lambda x}} +$

$\tilde{p}_B \frac{e^\lambda}{1+e^\lambda} = \frac{e^\lambda}{1+e^\lambda} + \tilde{p}_A \left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)$ is the expected payoff of action D when the sender says that the actual payoff table is A, (c) $v(U|B) = (1-\tilde{p}_A)\frac{e^\lambda}{1+e^\lambda} + x(1-\tilde{p}_B)\frac{e^{\lambda x}}{1+e^{\lambda x}} = x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \tilde{p}_B \left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)$ represents the expected payoff of action A upon message B and (d) $v(D|B) = x(1-\tilde{p}_A)\frac{e^{\lambda x}}{1+e^{\lambda x}} + \frac{e^\lambda}{1+e^\lambda}(1-\tilde{p}_B) = x\frac{e^{\lambda x}}{1+e^{\lambda x}} - \tilde{p}_A \left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)$ denotes the expected payoff of sending message B in state B. We can then verify that

$$\tilde{q}_A(\tilde{p}_A, \tilde{p}_B) = \frac{1}{1+e^{\lambda\left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)(\tilde{p}_A - \tilde{p}_B)}}$$

and

$$\tilde{q}_B(\tilde{p}_A, \tilde{p}_B) = \frac{1}{1+e^{\lambda\left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)(\tilde{p}_B - \tilde{p}_A)}}.$$

Up to now we have identified a system of four linear equations with the unknowns $\tilde{p}_A$, $\tilde{p}_B$, $\tilde{q}_A$, and $\tilde{q}_A$. Since $\tilde{p}_B = 1 - \tilde{p}_A$ and $\tilde{q}_B = 1 - \tilde{q}_A$, we can reduce the four equations to two:

$$\tilde{p}_A(\tilde{q}_A) = \frac{1}{1+e^{\lambda\left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)(1-2\tilde{q}_A)}}$$

and

$$\tilde{q}_A(\tilde{p}_A) = \frac{1}{1+e^{\lambda\left(x \frac{e^{\lambda x}}{1+e^{\lambda x}} - \frac{e^\lambda}{1+e^\lambda} \right)(2\tilde{p}_A - 1)}}.$$

It is easy to see that $\tilde{p}_A^* = \tilde{q}_A^* = \frac{1}{2}$ is the only solution of the two equations above. It follows immediately that $\tilde{p}_B^* = \tilde{q}_B^* = \frac{1}{2}$ and $\tilde{\mu}_A^* = \tilde{\mu}_B^* = \frac{1}{2}$. Finally, we can calculate that $l_p(\tilde{p}_A, \tilde{p}_B) = t_p(\tilde{p}_A, \tilde{p}_B) = \frac{1}{2}$. □

Appendix D: Instructions of the Punishment Game

Welcome

Thank you for coming. The purpose of this session is to study how people make decisions in a particular situation. If you have any questions, feel free to raise your hand and your question will be answered so everyone can hear. From now until the end of the session unauthorized communication of any nature with any other participant is prohibited. The experiment will be conducted through computers and all interactions between you will take place through them.

During the session you will play a game that gives you the opportunity to make money. What you earn depends partly on your decisions and partly on the decisions of others. At the end of the session, the amount you earned will be paid to you privately in cash.

We start with a brief instruction period. During the instruction period you will be given a description of the experiment. We are about to begin.

General Instructions

In your envelope you will find a questionnaire and an official receipt. Fill in the questionnaire and write down your name and matriculation number in the receipt. You will need both forms to receive your payment at the end of the session. Your personal data will be kept confidential and will be used for statistical purposes only.

In this session, you will play a game which is repeated for 50 rounds. Before the first round, the computer will randomly divide the participants into groups of six. This division will last for the entire session. Participants within each group will play only among themselves. The assignment process is random and anonymous so you will not know who is in your group.

Next, we will go over a brief tutorial. Please interrupt at any time if you have a question.

At the beginning of each round, you will be randomly joined with another participant from your group to form a pair. In each pair, one participant is randomly chosen to be the **Sender**, and one to be the **Receiver**. Remember that this process is random and the assignment changes every round.

Each round, after pairs have been formed and roles have been assigned, the computer selects one of the following two payoff tables. Final payoffs for both participants will be determined according to the selected table and the action **U** or **D** taken by the Receiver later on.

Table A	Sender	Receiver
Action U	2 Points	1 Point
Action D	1 Point	2 Points

Table B	Sender	Receiver
Action U	1 Point	2 Points
Action D	2 Points	1 Point

Sender's Instructions

At the beginning of the round **only** the Sender will be informed about the actual payoff table chosen by the computer. The Sender is the first one to take a decision in the game. S/He must communicate to the Receiver whether the payoff table chosen by the computer is either table **A** or table **B**. **Please, take into account that the Sender is free to tell the truth or to lie**. The computer screen for the Sender is as follows:

Put Figure 10 about here (caption: none). The corresponding graphic file is called figure3.eps and should be scaled down to 80%.

The two tables at the top of the screen represent payoffs according to tables **A** and **B**. Below you find the information whether table **A** or table **B** was chosen by the computer (in our example it is table **B**). On the inferior right corner there are two buttons labelled **A** and **B**. By clicking on the buttons **A** or **B** you inform the Receiver that you have observed the corresponding table. The Sender has 20 seconds to take this decision. **This is the only decision the Sender takes.**

Receiver's Instructions

The Receiver takes two decisions. First, once the Receiver got the Sender's message, s/he has to decide between actions **U** and **D**. The computer screen for the Receiver is as follows:

Put Figure 11 about here (caption: none). The corresponding graphic file is called figure4.eps and should be scaled down to 82.7%.

The two tables at the top of the screen represent payoffs according to tables **A** and **B**. Below you find the message from the Sender regarding the table s/he observed (in our example the Sender has informed the Receiver that s/he observed table **A**). On the inferior right corner there are two buttons labelled **U** and **D**. By clicking on the buttons **U** or **D** you take the corresponding action. The Receiver has 20 seconds to take this decision. Once this action is taken, a new screen appears summarizing the outcome of the round so far.

Put Figure 12 about here (caption: none). The corresponding graphic file is called figure5.eps and should be scaled down to 74.5%.

Now the Receiver is asked to take the second decision: S/He must either accept the current payoff distribution or reduce the payoff of both participants to zero. By clicking on the button **Reduce Payoffs** or **Accept Payoffs**, the Receiver takes the corresponding action. The Receiver has 15 seconds to take this decision.

Summary of the Round

The final screen is a summary of the round: It indicates the actual payoff table, the message chosen by the Sender, the actions taken by the Receiver, and the earnings of both participants in this round. Additionally, you are also informed about your accumulated payoff.

Put Figure 13 about here (caption: none). The corresponding graphic file is called figure6.eps and should be scaled down to 68%.

The screen above is the Receiver's summary. It indicates that the Sender chose message **A** whereas the Receiver took action **D** and accepted the payoffs. Therefore, the Sender gets 2 Points and the Receiver 1 Point. At the end of a round, click on **Continue**. The experiment will nevertheless proceed automatically to the next round in 10 seconds.

Payment

The Points you accumulate during the course of the session will determine your payment in addition to the £**5** show-up fee. The exchange rate Points/£ is **10p per Point**. At the end of the experiment, take your questionnaire and receipt to the counter for payment. They will be matched to our computer printout. Once you are paid, you may leave.

Figures and Tables

Table A	Sender	Receiver
Action U	x	1
Action D	1	x

Table B	Sender	Receiver
Action U	1	x
Action D	x	1

Table 1: Payoff Tables

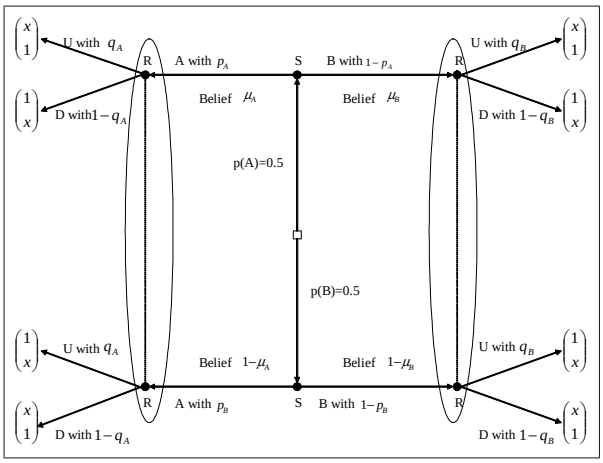
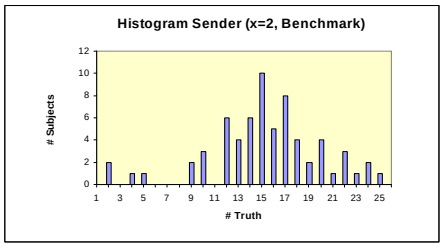


Figure 1: The Benchmark Game



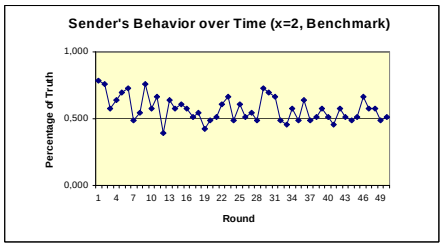
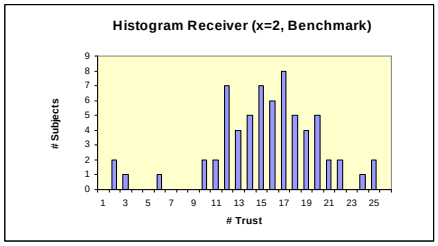


Figure 2: Senders' Behavior (x=2, Benchmark)



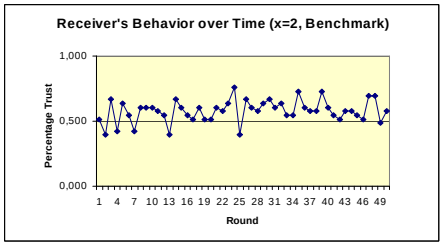
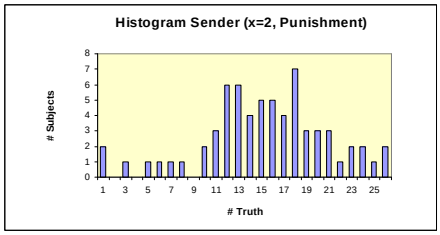


Figure 3: Receivers' Behavior (x=2, Benchmark)

	Truth	Lie
Trust	0	0.2525
Distrust	0.0546	0.0163

Table 2: Punishment Behavior ($x = 2$)



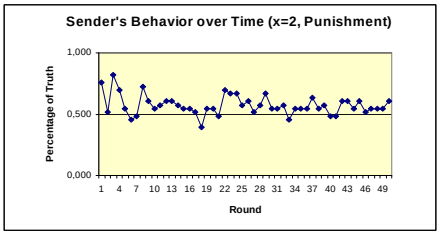
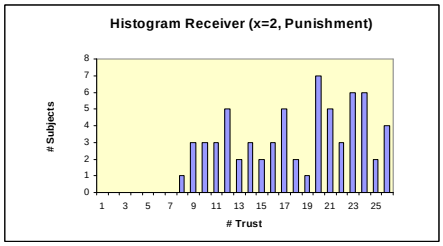


Figure 4: Senders' Behavior (x=2, Punishment)



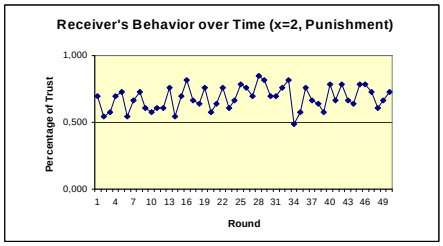
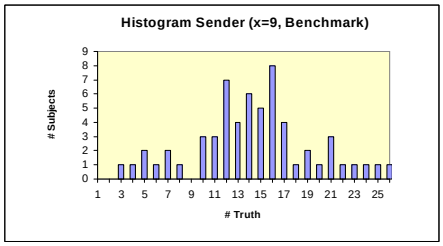


Figure 5: Receivers' Behavior (x=2, Punishment)



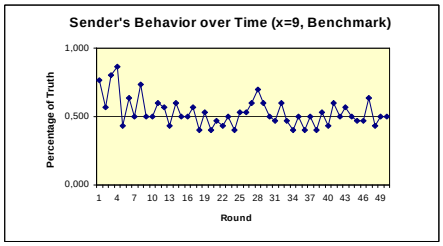
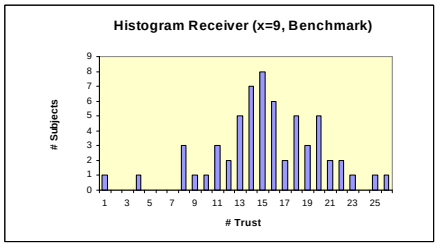


Figure 6: Senders' Behavior (x=9, Benchmark)



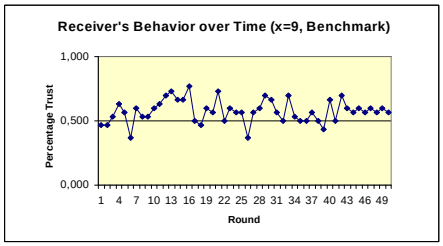


Figure 7: Receivers' Behavior (x=9, Benchmark)

	Truth	Lie
Trust	0	0.4286
Distrust	0.1340	0

Table 3: Punishment Behavior ($x = 9$)

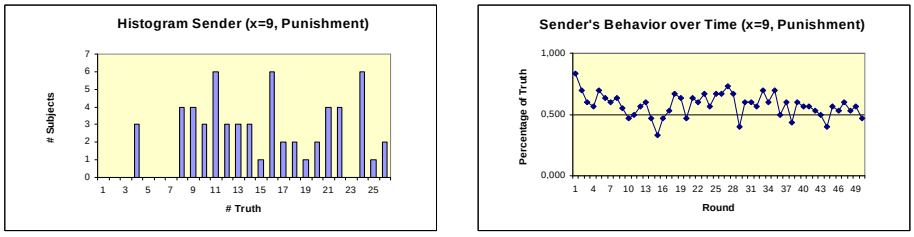


Figure 8: Senders' Behavior (x=9, Punishment)

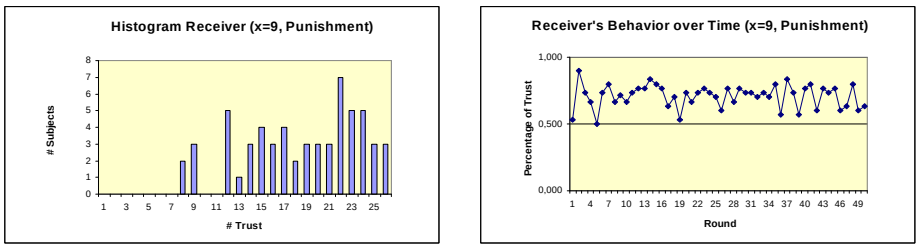


Figure 9: Receiver's Behavior (x=9, Punishment)

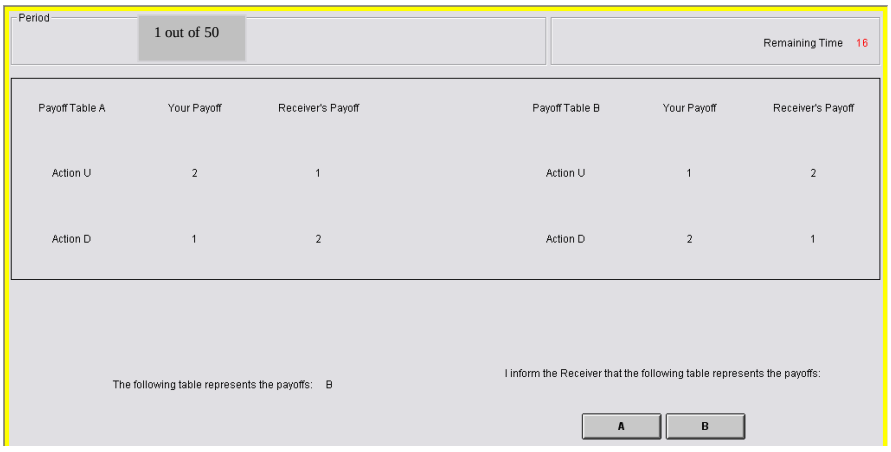


Figure 10: without caption

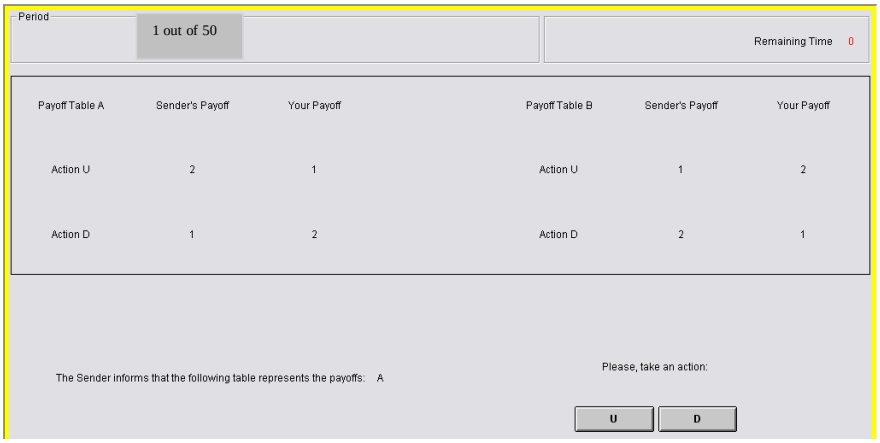


Figure 11: without caption

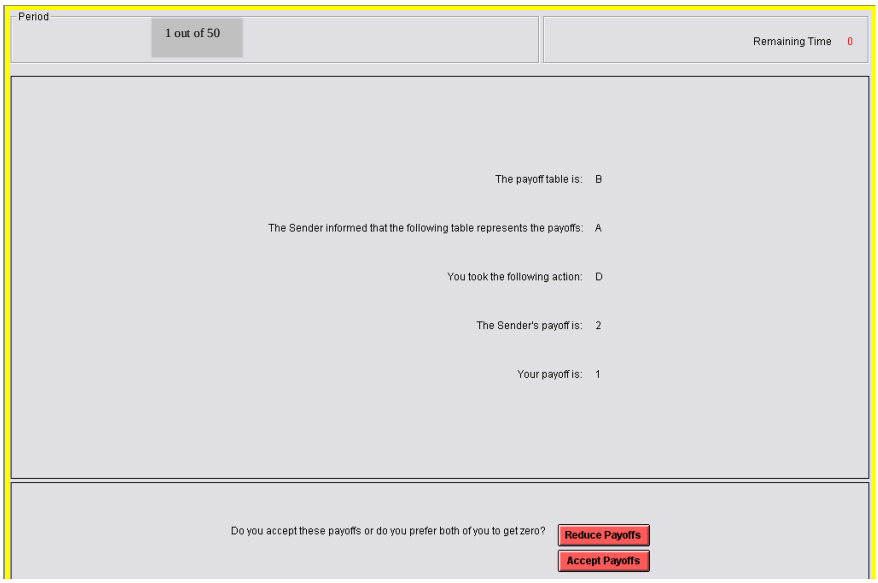


Figure 12: without caption

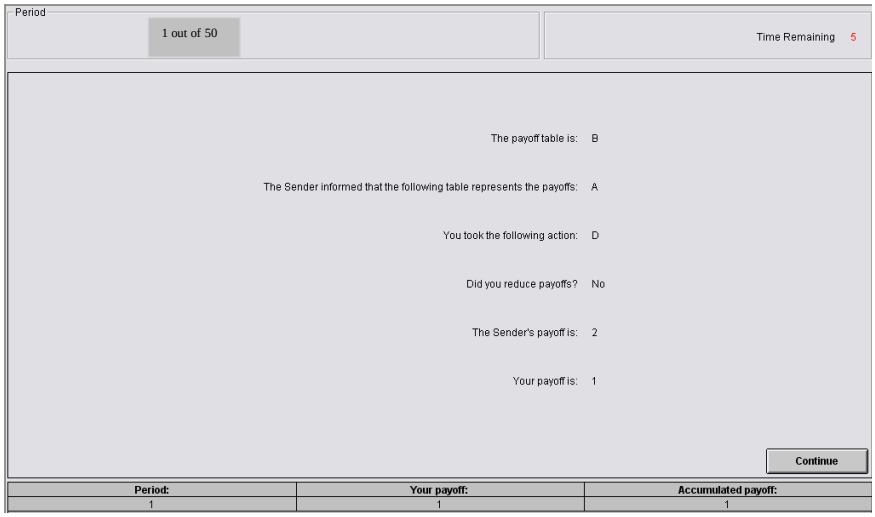


Figure 13: without caption